BUSA8000 – TECHNIQUES IN BUSINESS ANALYTICS – Group 78

No.	Name	Student ID
1	Sakshi Dilip Channawar	48078581
2	Darin Engkawong	48198889
3	Thi Hieu Ngan Tran (Rosa)	47747935
4	Mai Trong Nghia Hoang (Edward)	48728705

USE OF RESOURCES AND TECHNOLOGIES INCLUDING GENERATIVE ARTIFICIAL INTELLIGENCE

For this assessment, students are permitted to use generative artificial intelligence tools (GAITs e.g., ChatGPT) to:

- clarify concepts, theories, ideas, etc., discussed in class
- generate preliminary ideas for writing and coding
- edit a working draft of the assessment
- read and summarise research and supporting evidence for the assessment

Students are not permitted to use GAITs to

- Generate definitions or writing used in their final submission.
- produce counter-arguments or refine thinking on their final submission
- Generate complete Python code in their final submission.

Any of these actions will constitute and be treated as a breach of academic integrity.

Don’t’s

1. DON'T ask a GAIT to complete an assessment for you. This is outsourcing your assessment and is a breach of academic integrity.
2. DON'T blindly trust GAIT information. GAIT outputs can be completely inaccurate and will often contain fake references.
3. DON'T rely on GAITs to replace your own thinking and creativity.

Acknowledgement Statement by students:

Please select one acknowledgment from the following

☐ We acknowledge that we have not used GAITs (e.g., ChatGPT) in drafting and proofreading of this assignment.

☒ We acknowledge that we have only used GAITs (e.g., ChatGPT) in drafting and proofreading this assignment, which is permitted in the assignment instructions.

Executive Summary

This report analyzes LuminaTech Lighting, focusing on sales performance, customer retention, and operational efficiency. The objective is to identify profitability drivers, predict future sales, and assess churn risks using predictive modelling and machine learning. Data cleaning addressed missing values, outliers, and inconsistencies, while exploratory analysis uncovered key sales patterns and customer behaviour, crucial for building predictive models.

Key Findings and Insights

Sales trends indicated that cost management and sales volume are vital to profitability, with certain product categories significantly contributing to revenue. Customer retention and operational efficiency were influenced by factors like purchase frequency, refund behaviour, and engagement, highlighting areas for improvement. Predictive models using Multiple Linear Regression (MLR) and Random Forest (RF) were applied to forecast sales and predict churn, with the RF model outperforming MLR, especially in sales forecasting. Churn prediction faced challenges with class imbalance, which was effectively addressed via under sampling.

Challenges Encountered by LuminaTech

- Over-reliance on "make-to-order" products increases costs and churns.
- Seasonality impacts revenue, requiring better demand forecasts.
- High products' refund/return rates correlate with churn, indicating product issues.
- Delays in delivering orders (time gap) signal inefficiencies in production and logistics.

Strategic Recommendations

- Focus on suggested high-margin environmental products and optimize inventory.
- Expand the product portfolio, improve quality, and reduce returns.
- Streamline deliveries and implement a seasonal lighting strategy.
- Introduce a loyalty program and targeted promotions.
- Leverage sales prediction models to optimize operations.

This report equips LuminaTech with insights to optimise sales, reduce churn, and streamline operations. The recommendations provide strategic action for strengthening business performance through targeted, data-driven interventions.

I. Introduction

This report presents an analysis conducted for LuminaTech Lighting ("LuminaTech") to drive data-informed strategies in sales, customer retention, and operational improvement. The analysis begins with data cleansing to ensure a reliable dataset. Key insights are then extracted through visualizations, revealing trends in sales patterns, inventory performance, and customer demographics. This is followed by two sample tests to compare data subsets and multiple regression analysis to identify factors impacting business performance. Based on these findings, a predictive model is developed to forecast 2014 sales with optimized accuracy, and, finally, a likelihood model identifies customer segments at higher risk of churn. This structured approach provides LuminaTech's management team with precise recommendations to enhance sales strategies, deepen customer engagement, and boost business performance through targeted, data-driven insights.

II. Data Cleaning Process

The data cleaning process is crucial in transforming raw data into a format ready for analysis. For this project, several key procedures were systematically undertaken to enhance data accuracy, remove redundancies, and ensure consistency. The following sections detail each step, covering the rationale, methodology, and specific choices made throughout the process.

1. Initial Data Assessment

Prior to detailed cleaning, an initial review of the dataset's structure and completeness was performed using summary functions. This initial overview helped identify basic issues, such as missing values, duplicate entries, and potential data type inconsistency (Granti and Sarma, 2013). The following sections detail each step, covering rationale, methodology, and specific choices made throughout the process.

2. Handling Duplicates

Removing duplicates of 8,962 rows is critical to maintaining data integrity and avoiding inflated results, especially in metrics like sales totals, customer counts, or transaction volumes. Duplicate records can arise from data entry errors, system lags, or other operational inconsistencies. If not addressed, these duplicates could skew analysis results, leading to inaccurate insights and misleading conclusions. *Invoice_number, line_number, customer_code, item_code, invoice_date,* and *value_sales* were key indicators to remove duplicates to ensure each transaction's unique characteristics were captured. For instance, distinguishing products within the same invoice (line items) prevents mistakenly removing legitimate entries that could appear similar but are indeed unique. Keeping only the first occurrence of a duplicate retains the most complete version of each transaction while

eliminating redundant entries. Post-removal validation confirms that the dataset's shape aligns with expectations, verifying that only true duplicates were removed.

3. Handling Missing Values and Irrelevant Columns

a) Identifying and Treating Missing Values

Missing values, particularly in crucial columns, can reduce the reliability of the dataset and potentially cause errors in statistical analysis or modelling. Addressing these values ensures that each analysis can proceed without gaps or biases, leading to more robust conclusions. The column of *Item_source_class* (100% null) was removed since its complete lack of data means it cannot contribute to subsequent analysis. Retaining columns with 100% null values is not only redundant but could also create unnecessary complexity.

b) Removing Irrelevant or Redundant Columns

Not all columns in a dataset contribute value to analyses. Some may be operationally specific, redundant, or not aligned with the project's objectives. Removing such columns enhances focus, reduces computational load, and prevents unnecessary noise in the data. In this case, *dss_update_time* is system-specific timestamp, which does not influence transaction insights and would only clutter the dataset.

Additionally, given the presence of *invoice_date* and *order_date*, which contain full date information, *calendar_day* is redundant. Reducing these redundancies avoids overlapping information and streamline analyses involving date-based insights. However, *fiscal_year* and *calendar_year* were retained, as these fields allow for both operational and financial perspectives, which may reveal insights unique to each dimension. Retaining both offers flexibility for analyses involving temporal segmentation.

Removing columns with missing values, eliminating redundancies, correcting misspelled entries, and filtering undocumented or inactive codes, significantly enhanced the data's accuracy and quality, ensuring more reliable insights for downstream analysis (Komorowski et al., 2016).

4. Handling Inconsistencies in Categorical and Date Values

Consistent data types are fundamental for accurate data manipulation, analysis, and interpretation (Sharma, 2013). For instance, date fields must be in a datetime format to allow for chronological operations, such as time-based aggregations or interval calculations. String fields should be standardised to avoid misinterpretation due to extraneous spaces or mixed

formats. This means removing non-standardised or undocumented codes ensures cleaner and more interpretable results.

a) Data Type Standardisation

Converting date fields like *accounting_date, invoice_date*, and *order_date* from integer to datetime format enables us to accurately measure time intervals and perform date-based filtering. Regarding categorical fields*, customer_code* and *item_code* should be represented as strings, without leading or trailing spaces to helps maintain consistency. Any inconsistency in data representation could **cause issues in grouping, filtering, or analysis steps,** especially in categorical features essential for segmentation.

b) Correcting Categorical Inconsistencies

Standardising categorical data ensures that each category is represented consistently. For instance, minor variations in naming can create duplicate entries that disrupt aggregations and distort results. In *order_type_code,* replacing **PME** with **PMO** addressed potential misspellings based on frequency and metadata guidance. Without these corrections, analysis involving order types would yield misleading segmentation results. Moreover, rare codes in *warehouse_code* not documented in the metadata (e.g., **1N2** and **1N3**) were removed due to low frequency and lack of context.

c) Handling Currency and Logical Inconsistencies

Logical inconsistencies were also identified in the dataset, specifically with respect to the relationship between "*order_date"* and *"invoice_date*". A review revealed 22 instances where invoice dates predated their corresponding order dates, which is not feasible in practical business scenarios. As a result, these rows were dropped to preserve the logical flow of the data. Moreover, accurate and consistent currency representation is critical when dealing with financial data across multiple currencies. Currency standardisation was performed to address discrepancies in currency codes. Moreover, incorrect entries, such as **AUS** instead of **AUD**, were corrected. The currency column was also stripped of any leading or trailing spaces, ensuring uniformity.

Since the company operates primarily in Australia (AUD is the reporting currency), conversion was applied to all non-AUD transactions (e.g., NZD, USD, EUR) using the average exchange rates obtained from S&P Capital IQ for the corresponding periods (Appendix.4). This conversion enabled direct comparability in monetary values, particularly for fields like *value_sales* and *value_cost*, across various currencies, allowing for straightforward comparisons, aggregations and modelling without currency-related distortions.

5. Handling Outliers and Negative Values

Outliers, especially in financial data, can skew results, making averages, trends, and other statistics less representative of the dataset's core behaviour. Removing extreme values (based on the IQR) prevents these points from disproportionately influencing outcomes, especially in sales, cost, and quantity analyses. Key numeric fields (*value_sales, value_cost,* and *value_quantity*) were assessed for outliers using the Interquartile Range (IQR) method. Outliers beyond the 1.5 x IQR bounds were removed to reduce skewness and prevent extreme values from disproportionately influencing analysis results, focusing on the central trend in these variables.

Negative values in sales, costs, or quantities required careful handling. While negative values can legitimately represent returns or refunds, they must be correctly classified to prevent confusion with normal sales transactions. Incorrect handling of negatives could lead to inaccurate financial assessments or forecasts. In this project, negative sales values initially coded as **NOR** (Normal Orders) were reclassified to **CRR** (Credit/Return). This prevented mixing returns with regular sales, which would distort sales performance metrics. Nevertheless, negative values not associated with refund codes (e.g., CRR, COP) were removed, as they likely represent data entry errors rather than legitimate returns. This cleanup ensures that only accurate, meaningful data remains.

III. Exploratory Insights

This section analyses the business overview, performance trend, and extract interesting insights for LuminaTech management team. However, the negative values e.g. product return, credit price adjustments are not considered as outliers and are not excluded from the dataset since they reflect actual business performance and can lead to interesting insights on both business and operational improvement e.g. large loss from product return, or price adjustment. Removing these negative values could obscure these potentially valuable signals that highlight unique patterns, customer behaviors, or operational challenges.

1. Sales Trend Analysis

a) Approach

The monthly sales trends were analysed using line charts that differentiate between project-based and non-project-based sales. Two distinct lines, one solid for non-project-based sales and one dashed for project-based sales, were color-coded by calendar year to allow performance comparison between 2012 and 2013 (Figure.1).

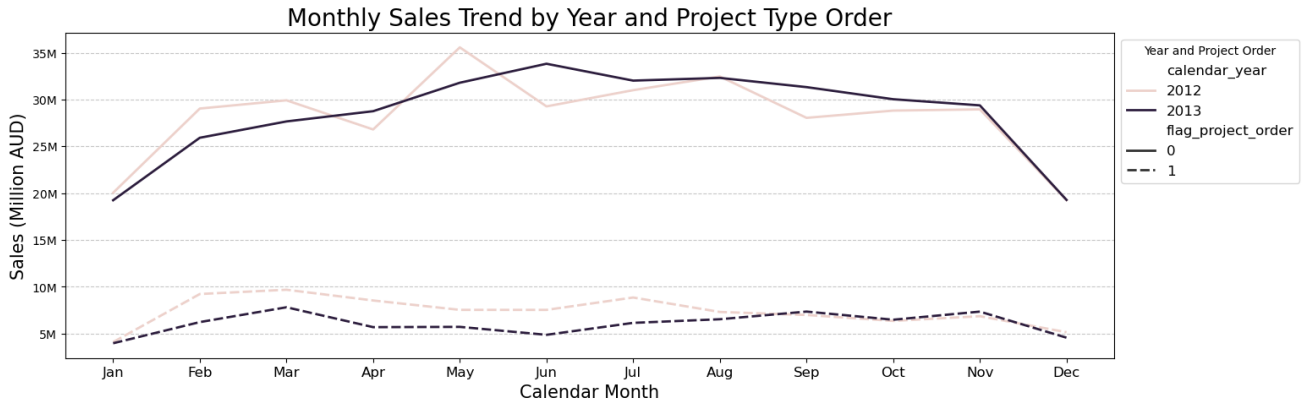

Figure.1: Monthly Sales Trend by Year and Project Type Order

b) Insights

Firstly, LuminaTech recorded a total revenue of AUD 403.5 million in 2013, marking a 3% decline from 2012. This reduction was primarily attributed to a 16% year-on-year (YoY) decrease in project-based sales, which constituted 20% of total sales, while non-project-based sales (80% of total) remained stable with a slight increase of 0.3% YoY (Figure.2). Secondly, the lowest sales were observed in December and January, likely indicating a seasonal dip in demand during the Australian summer holiday months. A portion of the company's revenue was in foreign currencies.

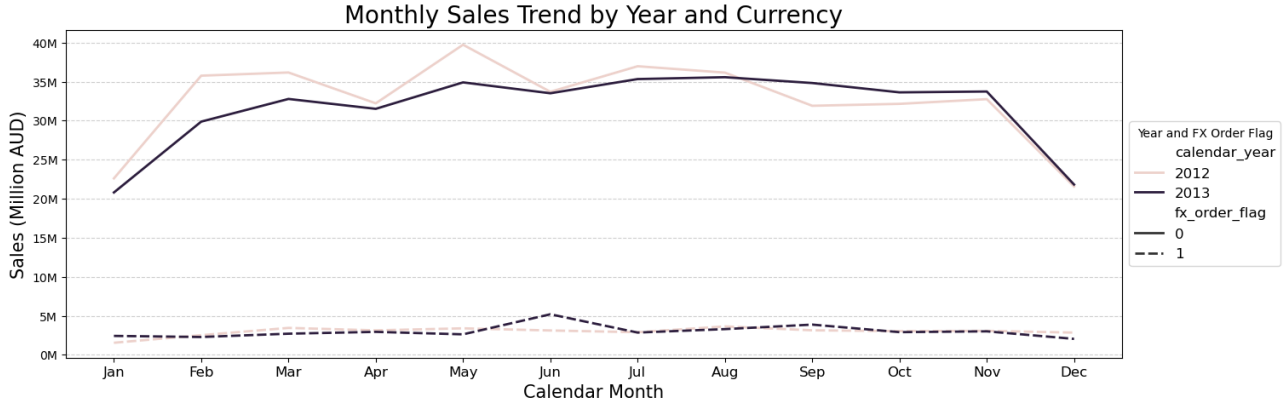

Figure.2: Monthly Sales Trend by Year and Currency

Lastly, unlike the AUD revenue, which decreased in 2013 due to project-based sales, foreign currency revenue remained stable, suggesting that the drop in project-based sales was primarily a domestic issue. The company also has a small portion of foreign currency revenue. The foreign currency 2013's performance was constant, while AUD revenue dropped in 2013 from project-based sales.

c) Actionable Recommendations for Management Team

The decline in project-based sales calls for further investigation, specifically regarding **potential customer delays** or contract losses from project-based orders. Moreover,

recognizing seasonality also enables better inventory and production planning, aligning resources with periods of high and low demand.

2. Profitability Across Customer Districts

a) Approach

Profit margin by customer district was calculated as (Value Sales/Value Cost-1). This included adjustments for returns, refunds, and price changes, providing a realistic view of profitability.

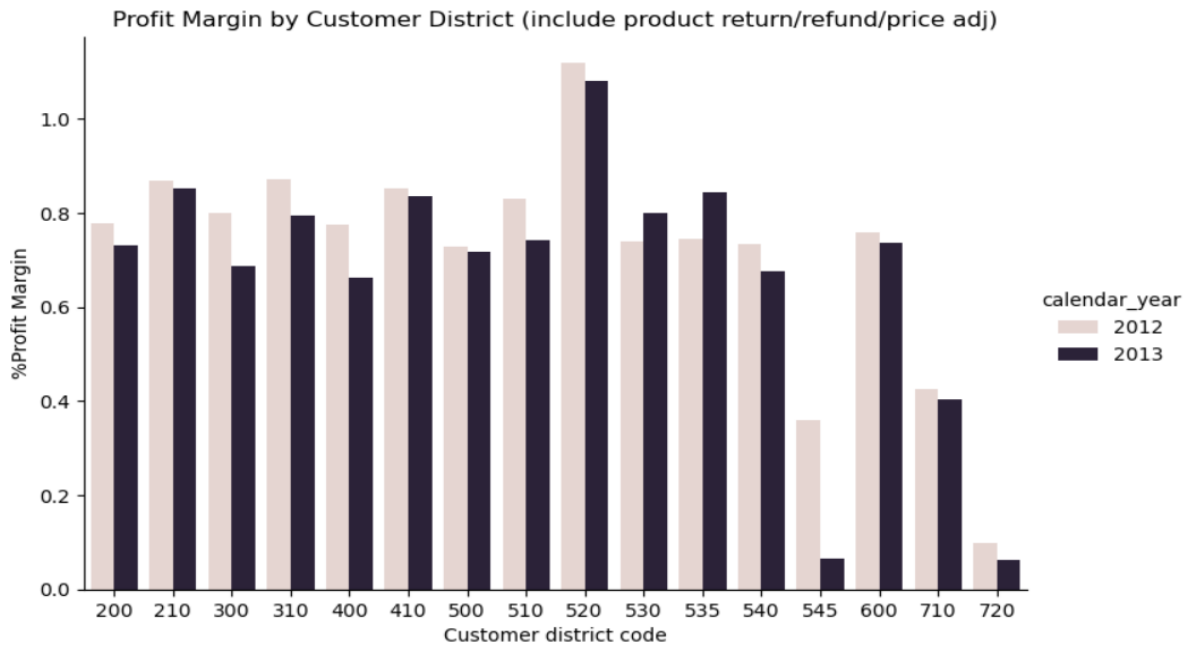

Figure.3: Profit Margin by Customer District

b) Insights

Low Profitability in Intercompany and Head Office Transactions

Intercompany (720), Head Office (710), and Head Office NZ (545) had notably low margins due to their roles in intra-group transactions. The margin for Head Office NZ (545) dropped sharply in 2013, possibly due to strategic profit allocation.

Profit Margin Decline in Key Districts

Major customer districts, including Melbourne (the second-highest sales district), experienced significant margin declines in 2013. Only South (530) and Central (535) island NZ showed improved margins, albeit with limited impact on overall sales.

c) Management Focus

Profit margin is a critical metric, particularly for high-impact districts like **Melbourne (300).** Management should investigate causes of margin reductions, which may include shifts in

product mix toward lower-margin items. Monitoring margin trends across districts helps identify inefficiencies in cost control and product preference shifts.

3. Profit Margin and Currency

a) Approach

Profit margin by currency was calculated in the same way as district margins. This aggregation at the currency level provided insight into the profitability of sales in AUD, NZD, USD, and other currencies.

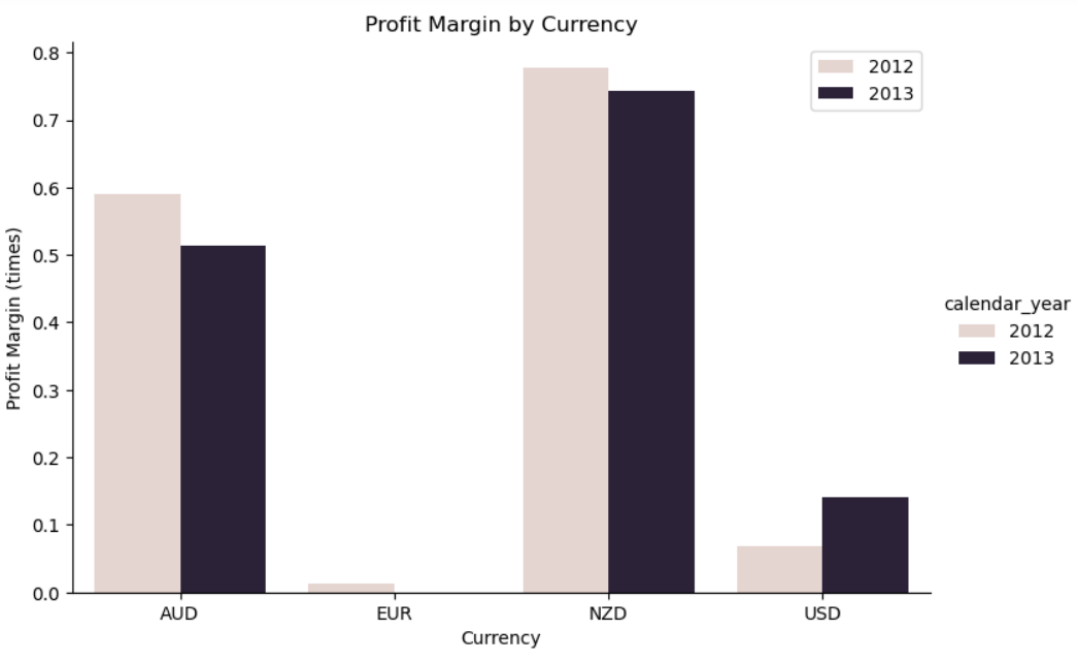

Figure.4: Profit Margin by Currency

b) Insights

Although AUD is the primary currency, its margins were lower than NZD. This disparity was due to the low-margin intercompany and head office transactions in AUD. However, the low margin does not necessarily indicate competitive pressure but reflects **potential issues in strategic pricing** within the group.

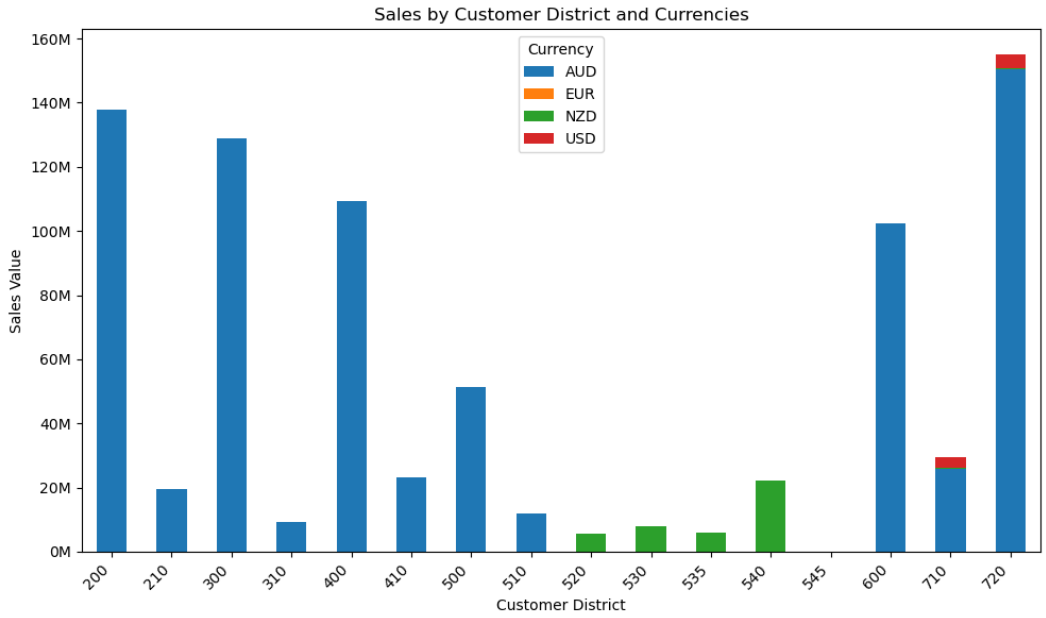

Figure.5: Sales by Customer District and Currencies

Revenue in USD, on the other hand, was limited to head office (710) and intercompany (720) sales, resulting in low margins.

4. Order-to-Invoice Time Gap Analysis

a) Approach

The analysis measures the time gap between the order date and the invoice date to assess how quickly LuminaTech processes orders and delivers products to customers. This is visualized through a histogram of time gap categories, which highlights the distribution of delivery times across different types of orders.

b) Insights

Most transactions were completed within a time gap of **0-3 days**, indicating efficient handling for a large portion of orders. This suggests that LuminaTech's typical operations meet delivery expectations promptly, especially for standard, non-project orders. However, the chart indicates that a certain portion of transactions took over **30 days**. The extended delivery times primarily **stem from project-based orders**, which often have more complex requirements and specifications, potentially resulting in longer delivery times.

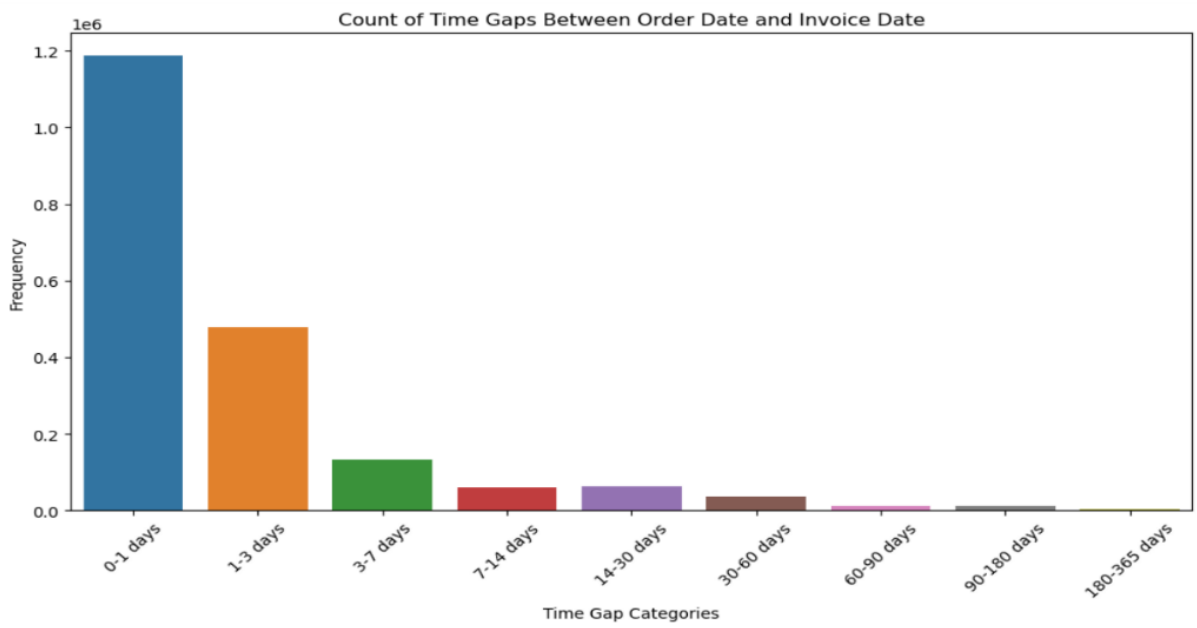

Figure.6: Count of Time Gaps Between Order Date and Invoice Date

In non-project orders, certain product classes, particularly **Class D** (*low_volume & low_contribution*) and **Class I** (Indent), show a tendency for extended delivery times. This pattern suggests there could be **operational inefficiencies** or **supply chain issues** specifically affecting these product classes.

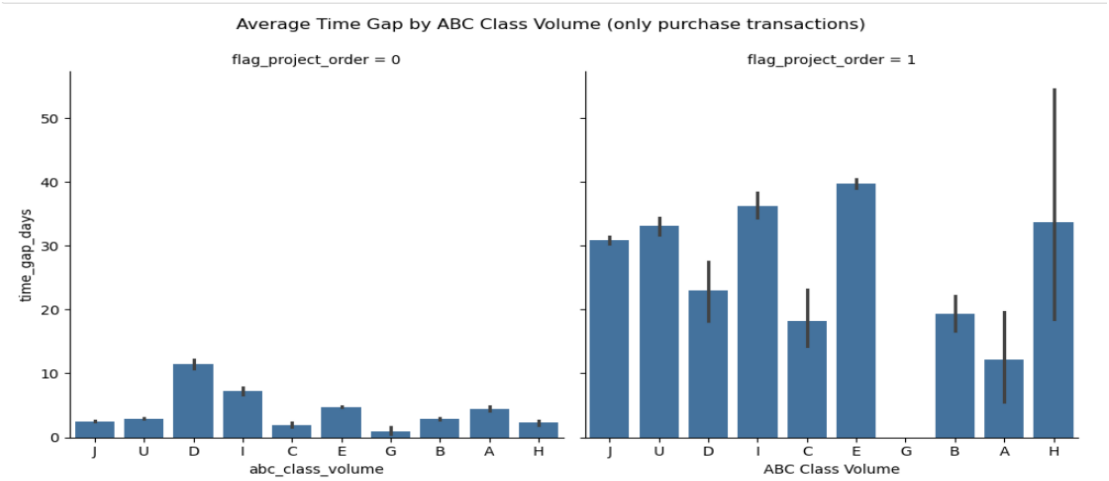

Figure.7: Average Time Gap by ABC Class Volume (only purchased transactions)

Finally, Class J, representing *make-to-order items*, constitutes over half of LuminaTech's sales. The consistent demand in this class presents an opportunity for inventory optimization. Therefore, by stocking certain high-demand products, LuminaTech could potentially reduce the time gap, enhance customer satisfaction, and prevent customer churn.

c) Actionable Recommendations for Management Team

This analysis enables LuminaTech to identify specific product classes or order types that require attention to improve operational efficiency. Addressing these time gaps can lead to better inventory management and customer satisfaction. For project-based orders, investigating ways to streamline the supply chain or production process could improve delivery timelines.

5. Customer Retention Rate

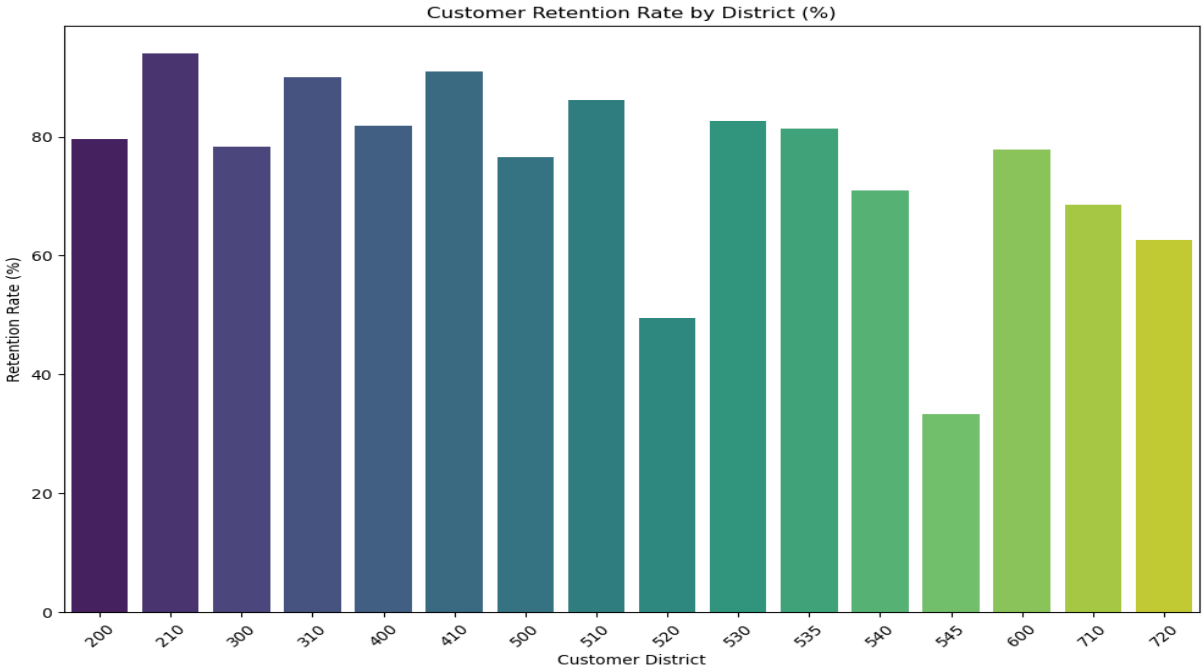

Figure.8: Customer Retention Rate by District (%)

a) Approach

The customer retention rate is calculated based on customers with orders placed after **30 June 2013,** which is around 6 months from the last order date (31 December 2013). Retention is analysed at the district level to understand customer loyalty across different geographical segments.

b) Insights

LuminaTech achieved a **78%** overall retention rate. This solid retention rate reflects a generally loyal customer base, though variances exist across districts. However, District 520 (Inlite NZ) and District 545 (Head Office NZ) reported significantly lower retention rates compared to other regions. Although these districts contribute marginally to overall sales, Inlite NZ (520) generates high-profit margins, making customer retention there more critical. The lower retention in these districts may be a result of **competitive pressures**, **product dissatisfaction**, or **lack of targeted marketing**. Additionally, given the profitability of Inlite

NZ, the company should investigate underlying factors contributing to low retention and develop strategies to enhance customer loyalty. Nevertheless, the core domestic markets, such as **Sydney (200), Melbourne (300),** and **Brisbane (400),** maintained strong retention rates. These high-retention areas provide a stable revenue base for the firm, suggesting effective service and satisfaction in these regions.

IV. Test Sub Sample Differences

a) Business Rationale

Retaining customers is one of the key success factors to sustain business. Thus, it is important to examine what factors cause customers to end their relationship with a company. Preliminary investigation is conducted by applying a t-test to determine whether there are significant differences in indicators between churn and non-churn customers.

b) Approach (T-test 2 Independent Samples)

While T-Test cannot prove the correlation or causality that these indicators lead to customer churn, this preliminary test revealed that there are significant differences in indicators between two groups of customers which we can further examine with regression models.

c) Two independent groups

- *Churn Customer Definition*: Customers who have no transaction with the company in the last 6 months (last order is within 30 Jun 2013)
- *Non-Churn Customer Definition:* Customers who still have transactions with the company in the last 6 months (last order is after 30 Jun 2013)

d) Two selected indicators

The customers' satisfaction depends on the quality and timeliness of the product they receive. Thus, the following indicators are created to represent these business performances.

- *Percentage of Product Return/Sales*: Poor product quality (resulting in product returns) may be a factor contributing to customer dissatisfaction
- *Time Gap* between *Invoice date* and *Order date*: The delay of product delivery may be a factor contributing to customer dissatisfaction

e) Hypothesis Testing Questions

a) Is there a difference in the average percentage of product return costs between Churn Customers and Non-Churn Customers?"

- **H0:** Mean of % Product Return in Churn customer = Non-Churn Customer
- **H1:** Mean of % Product Return in Churn customer ≠ Non-Churn Customer

Approach

- Percentage of Product Return/Sales: Product Return Amount / Original Sales before product return[1]
- Aggregate data at the customer level for calculation and include only customers who have return transactions.
- Check Normal Distribution: There are **984 churned customers** and **3,496 non-churn customers**. The distribution of their Percentage of Product Return/Sales are as follows:

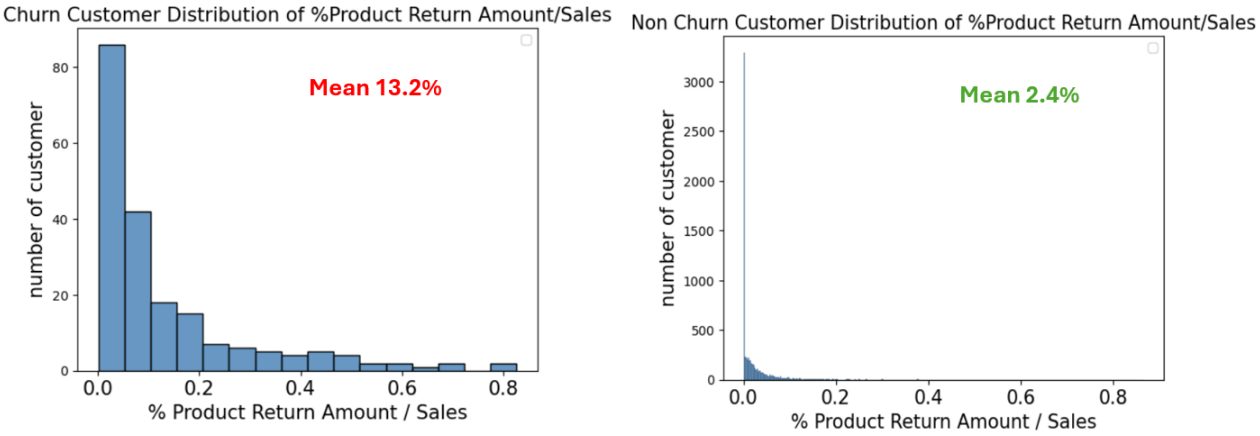

Figure.9: Churn and Non-churn Distribution of %Product Return Amount/ Sales

Interpretations

Percentage of product return (PPR) has a highly skewed right in both customer groups. This is a normal business situation in which most customers did not have high product return transactions. There are some customers with a high PPR of 50%-80% of sales which need further investigation to identify the root causes and develop a strategy to prevent this problem. With the above histogram and Shapiro-Wilk Test, % product return distribution in both customer groups is not normally distributed which is a required assumption for conducting T-Test. However, T-Test is still applicable as sample sizes are 984 for churn customers and 3,496 for non-churn customers which are large enough (over 30 samples) that the sampling distribution of the sample mean approaches normality according to Central Limit Theorem.

[1] Product Return Amount = Sum amount of product return transactions (order type code 'CDG', 'CRR', 'WDC', 'CPR', 'ZCG')

T-Test and Assumption Test results are shown in the below table

Test		Churn	Non-Churn	Conclusion
Shapiro-Wilk Test	Shapiro-Wilk Test	0.742	0.407	
	P-Value	0	0	Not Normally Distributed
Levene Test	Levene Test Statistic, P value	310 , 0.00		No Equal Variance
T-Test	Mean %product return/sales	13.25%	2.46%	
	T-Test	9.07		
	P-Value	0		Significanly Different Mean

T-Test rejects the null hypothesis meaning that there is a significant difference in %product return/sales between Churn customers (13%) and non-churn customers (2%) concluding that this indicator is significantly different among the two groups. However, it is not able to conclude that %product return causes or correlates with the customer's churn. It requires further tests with regression analysis.

b) Is there a difference in the average Time Gap between order data and invoice date between Churn Customers and Non-Churn Customers?

- H0: Mean of Time Gap in Churn customer = Non-Churn Customer
- H1: Mean of Time Gap in Churn customer ≠ Non-Churn Customer

Approach

- Time Gap is the difference between Order date and Invoice date.
- The test focuses on purchase transaction only (the refund, return, price adjustment transactions will be excluded from the tested data set)

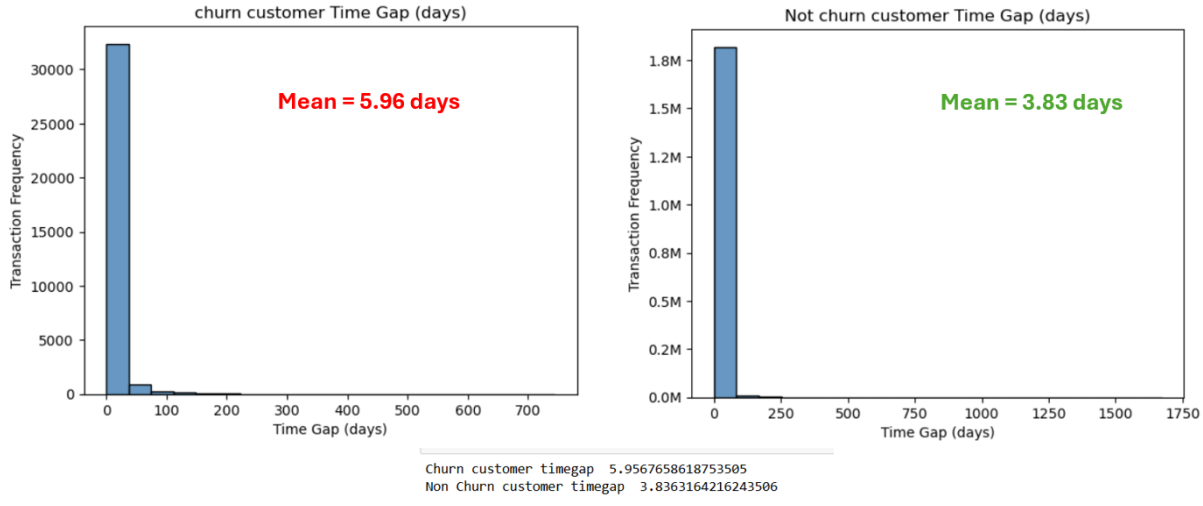

Figure.10: Churned and Non-churned Customers' Time Gap

As seen from the chart, time gap is highly right-skewed data since the project-type order that have long time gap accounts for only 20% of overall sales. Make-to-order inventory type results in an average time gap of over 0 days. There is also an abnormal long-time gap of over

365 days in both customer groups. The management should investigate the root causes of this time gap whether it was due to the company's inefficient operation, or it was due to the customers' project delay.

The Shapiro-Wilk Test is conducted for both Churned and Non-Churned customers' time gap distribution and found that they are not normally distributed. However, due to the large sample size, the central limit theorem can be applied in that the sampling distribution of the sample mean approaches normality; hence, t-test is still applicable.

Assumptions and Interpretations

Equal Variance: two customer group variances are not equal (Levene Test Statistic: 514, P-value 0.00). The T-test is conducted to examine mean of churn and non-churn customers on 2 datasets

a) all transactions including the unusually long-time gap; and
b) dataset includes only transactions with time gap not over 365 days.

Null hypothesis is rejected in both datasets meaning that there are significant differences in the time gap between Churn and Non-Churn Customers. The extremely high time gap transactions found in the above charts did not affect the mean and T-test results. The test results are as follows:

Value	All Transactions	Transactions with time gap in 365 days
Mean Churn Time Gap	5.95	5.86
Mean Non-Churn Time Gap	3.84	3.76
T-statistic	19.73	20.47
P-Value	0	0
Conclusion	Significant Difference between Churn, Non-Churn	Significant Difference between Churn, Non-Churn

f) <u>Insight Conclusion from two independent samples T-Test</u>

From both assumptions testing, there are significant differences in both %product return/sales and Time Gap between the order date and invoice date between the Churn customers and non-churn customers. Management should further investigate the root cause of these operational issues to determine whether they stem from the company's internal operations, third-party suppliers or freight services, or delays in product delivery due to customers' postponed projects. However, with T-test, it is unable to conclude that these two indicators cause or correlate with churn customers. Thus, these 2 indicators should be tested with a regression model.

V. Inference

Question 1: How do Customer Segments, Profit Margin, Cost, Quantity and Seasonal Trends Influence Value Sales?

Model 1	
y (Target Variable)	value_sale
x (Independent Variables)	value_cost, value_quanity, calendar_month_12, calendar_month_1, customer_district_code_530, customer_district_code_535, customer_district_code_520, customer_district_code_545.

These potential factors are integrated into the model as cost influences demand, higher prices can deter price-sensitive customers but may boost perceived quality. Sales volume reflects demand trends, while profit trends help in valuing sales patterns (Zisis et al., 2021). Regional factors also matter, with groups 530 and 535 showing the strongest performance in 2013, group 545 the lowest profitability, and group 520 maintaining relatively high profitability despite a slight decline (Section.2) (Fernandes, 2013). Seasonality also impacts demand, longer summer days reduce lighting needs, while shorter winter days increase them, particularly in December and January (Section.2) (Zisis et al., 2021). Together, these potentially improve sales valuation and strategic planning.

1. Multicollinearity

The three numerical independent variables in independent variable are *value_cost*, *value_quantity*, and *profit_margin*. Most VIF values are within the accepted range of -5 to 5, indicating low multicollinearity. VIFs exceeding 10 would signal a concern, but multicollinearity is not a major issue here.

Features	VIF
Const	2.752941
Value cost	1.005551
Value quantity	1.005489
Profit margin	1.000128

Table.1: VIF Checking for Model.1

2. Assumptions Checking

a. Linearity

The linearity assumption is assessed through the correlation coefficients between *value_sales* and the independent variables. *Value_cost* shows a strong positive correlation (0.8939), indicating a significant linear relationship. *Value_quantity* has a weak positive correlation (0.1052), while *profit_margin* (0.0013) and the calendar month variables show negligible correlations (Table.2). The customer district codes also have weak correlations. Nevertheless, the results support the linearity assumption, primarily due to the strong relationship with *value_cost*.

Value sales	1
Value cost	0.893853
Value quantity	0.105157
Profit Margin	0.001273
Calendar_month_1	-0.013897
Calendar_month_12	-0.000161
Customer_district_code_530	0.012972
Customer_district_code_535	0.002216
Customer_district_code_540	0.005547
Customer_district_code_545	-0.002053

Table.2: Linear Correlation Checking for Variables in Model.1

b) Independence & Homoscedasticity

While the homoscedasticity assumption stipulates that residuals should be randomly scattered around a horizontal line, the observed downward trend in the figure indicates a violation of this assumption. Additionally, the presence of trends or patterns in the residual plot suggests a lack of independence among the residuals and reinforces the violation of homoscedasticity.

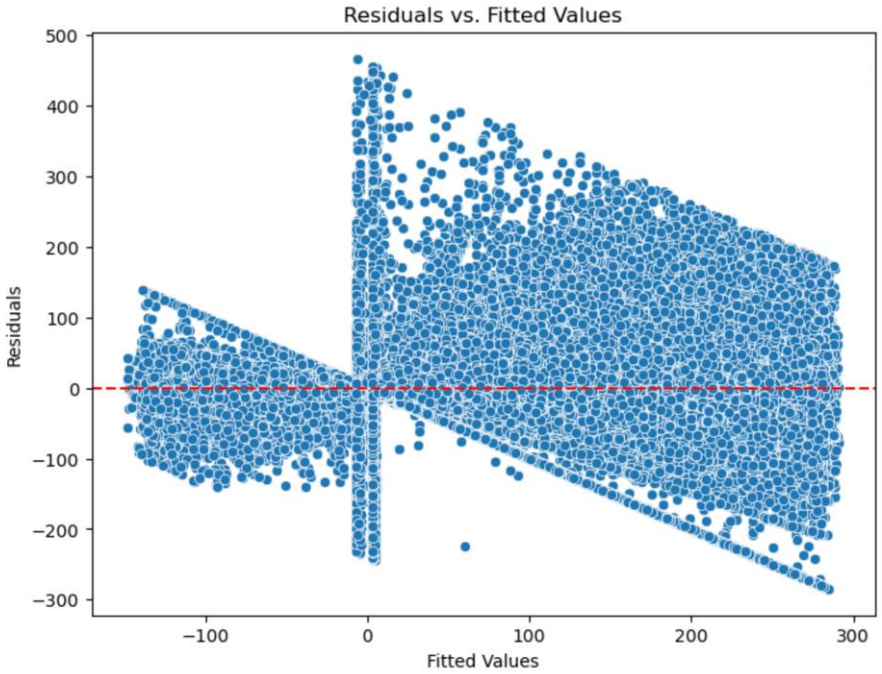

Figure.11: Residual vs Fitted Values of Model.1

c) Normality of Residuals

Since the residuals are approximately normally distributed, allowing for valid statistical inferences and reliable model predictions as the normality assumption is satisfied.

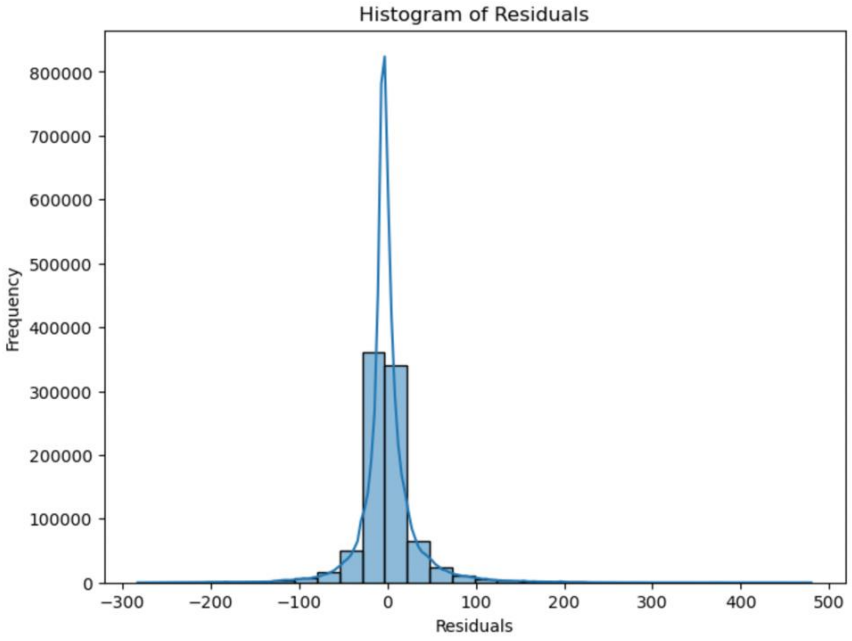

Figure.12: Histogram of Residuals

3. Output of Model 1

The OLS regression results show an R-squared of 0.801, indicating that 80.1% of the variance in sales revenue can be explained by the included independent variables, with a significant F-statistic of 399,900 (p<0.001). Key predictors include *value_cost* (1.8806, p<0.001) and *value_quantity* (0.4237, p<0.001), both significantly impacting sales revenue. The *profit_margin* (0.0003, p<0.001) and calendar_month_12 (1.6049, p < 0.001) also contribute positively. Significant negative effects are observed for *customer_district_code_530* (-7.8940, p < 0.001), *customer_district_code_535* (-9.6961, p<0.001), and *customer_district_code_520* (-37.2219, p<0.001), while *calendar_month_1* (-0.1810, p=0.170) and *customer_district_code_545* (-2.5399, p=0.739) are not significant.

Despite identifying these significant predictors, the model violates the independence and homoscedasticity assumptions (mentioned above). These issues suggest caution in interpreting the results, as they may affect the reliability of the estimates and predictions.

OLS Regression Results

Dep. Variable:	value_sales	R-squared:	0.802
Model:	OLS	Adj. R-squared:	0.802
Method:	Least Squares	F-statistic:	4.021e+05
Date:	Tue, 05 Nov 2024	Prob (F-statistic):	0.00
Time:	23:47:40	Log-Likelihood:	-4.3927e+06
No. Observations:	892625	AIC:	8.785e+06
Df Residuals:	892615	BIC:	8.786e+06
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.2310	0.060	69.991	0.000	4.113	4.349
value_cost	1.8830	0.001	1888.772	0.000	1.881	1.885
value_quantity	0.4326	0.005	86.502	0.000	0.423	0.442
profit_margin	0.0003	1.44e-05	21.858	0.000	0.000	0.000
calendar_month_1	-0.3626	0.132	-2.756	0.006	-0.620	-0.105
calendar_month_12	1.6742	0.162	10.358	0.000	1.357	1.991
customer_district_code_530	-8.4428	0.247	-34.191	0.000	-8.927	-7.959
customer_district_code_535	-10.2398	0.262	-39.112	0.000	-10.753	-9.727
customer_district_code_540	-11.1385	0.168	-66.247	0.000	-11.468	-10.809
customer_district_code_545	-2.9857	7.614	-0.392	0.695	-17.909	11.938

Omnibus:	339838.709	Durbin-Watson:	1.398
Prob(Omnibus):	0.000	Jarque-Bera (JB):	11431266.197
Skew:	1.190	Prob(JB):	0.00
Kurtosis:	20.369	Cond. No.	5.29e+05

Figure.13: Model.1's Output

Question 2: How do variations in cost, quantity, business area, customer district, and seasonal trends impact the brand-specific sales performance within environmental group S, P, J, and I?

Model 2 (for environmental groups – S and P)	
Y(i) (Target Variable)	value_salei
X(i) (Independent Variables)	value_costi, value_quanityi, j∑βj×customer_district_codej,i + k∑ βk×calendar_monthk,i + z∑βz×business_area_codez,i

Each factor, cost, quantity, customer district, business area, and seasonal trends, significantly influences brand-specific sales performance across environmental groups (S, and P). Products S and P exhibit strong seasonal trends, while P shows the highest profit margins and sales volumes across environmental product codes (Appendix.1). Cost and quantity are central to pricing strategy and demand, directly affecting revenue. The customer district accounts for geographical variations, capturing localised preferences and purchasing power, while the business area highlights differences in performance across operational segments. Seasonal trends captured by calendar month reflect fluctuating demand, influenced by holidays or weather, which varies throughout the year. Analysing these factors offers valuable insights into the drivers of profitability within these specific product brands.

1. Multicollinearity

The three numerical independent variables in independent variable are *value_cost, value_quantity*. Most VIF values are within the accepted range of -5 to 5, indicating low multicollinearity. VIFs exceeding 10 would signal a concern, but multicollinearity is not a major issue here.

Features	**VIF**
Const	2.752941
Value cost	1.005460
Value quantity	1.005460

Table.3: VIF Checking for Model.2

2. Assumption Checking

a) Linearity

The linearity assumption is evaluated through correlations between *value_sales* and independent variables. *value_cost* shows a strong positive correlation (0.901), supporting a linear relationship with *value_sales*. Conversely, *value_quantity* and *business_area_code_SUR* show weak positive correlations (0.152 and 0.110), and *business_area_code_LMP* a weak negative correlation (-0.239). Other variables, including

district codes, display negligible correlations (Appendix.2&3). Overall, these findings support the linearity assumption, mainly due to the strong correlation with *value_cost.*

b) Independence & Homoscedasticity

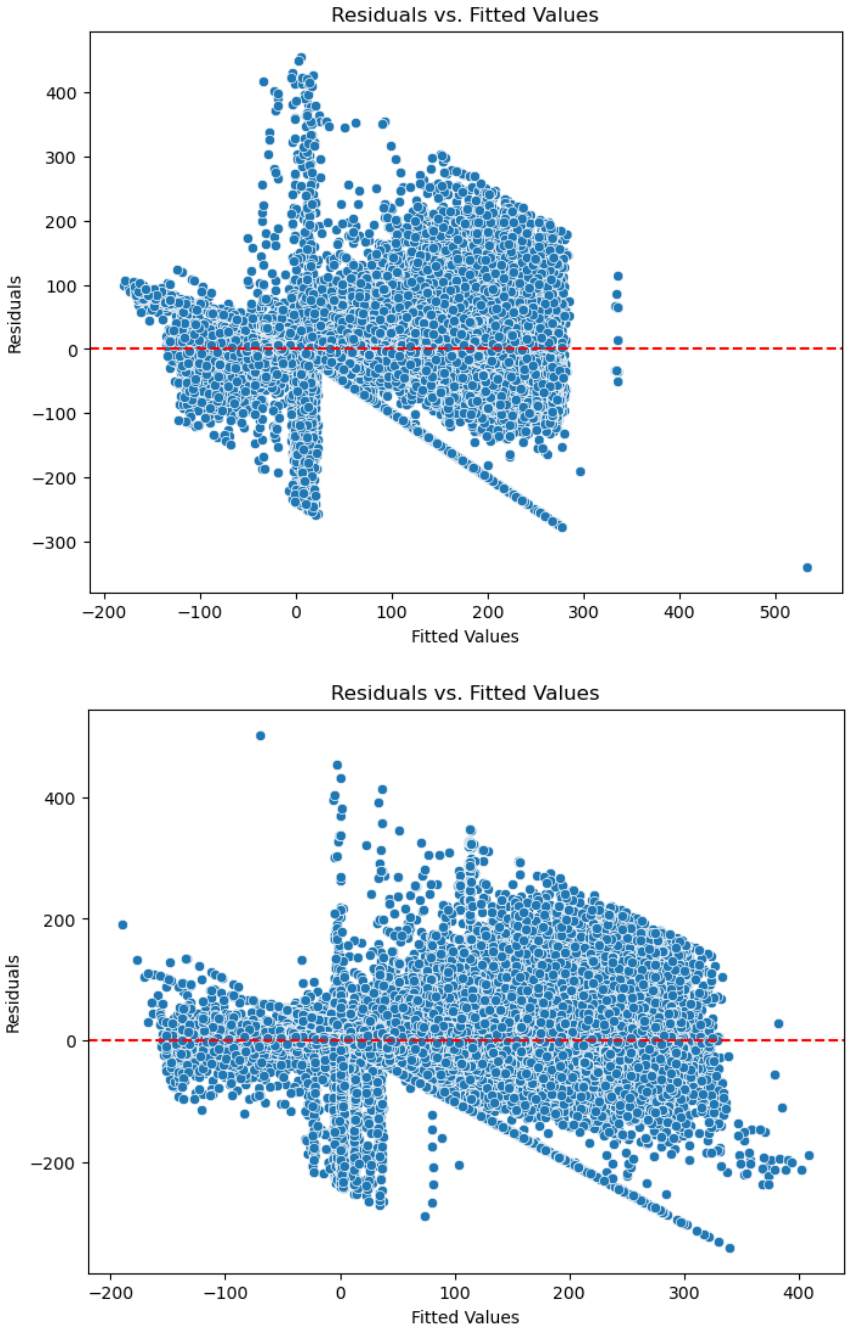

Figure.14. Residual and Fitted Values (S and P, respectively)

c) Normality of Residuals

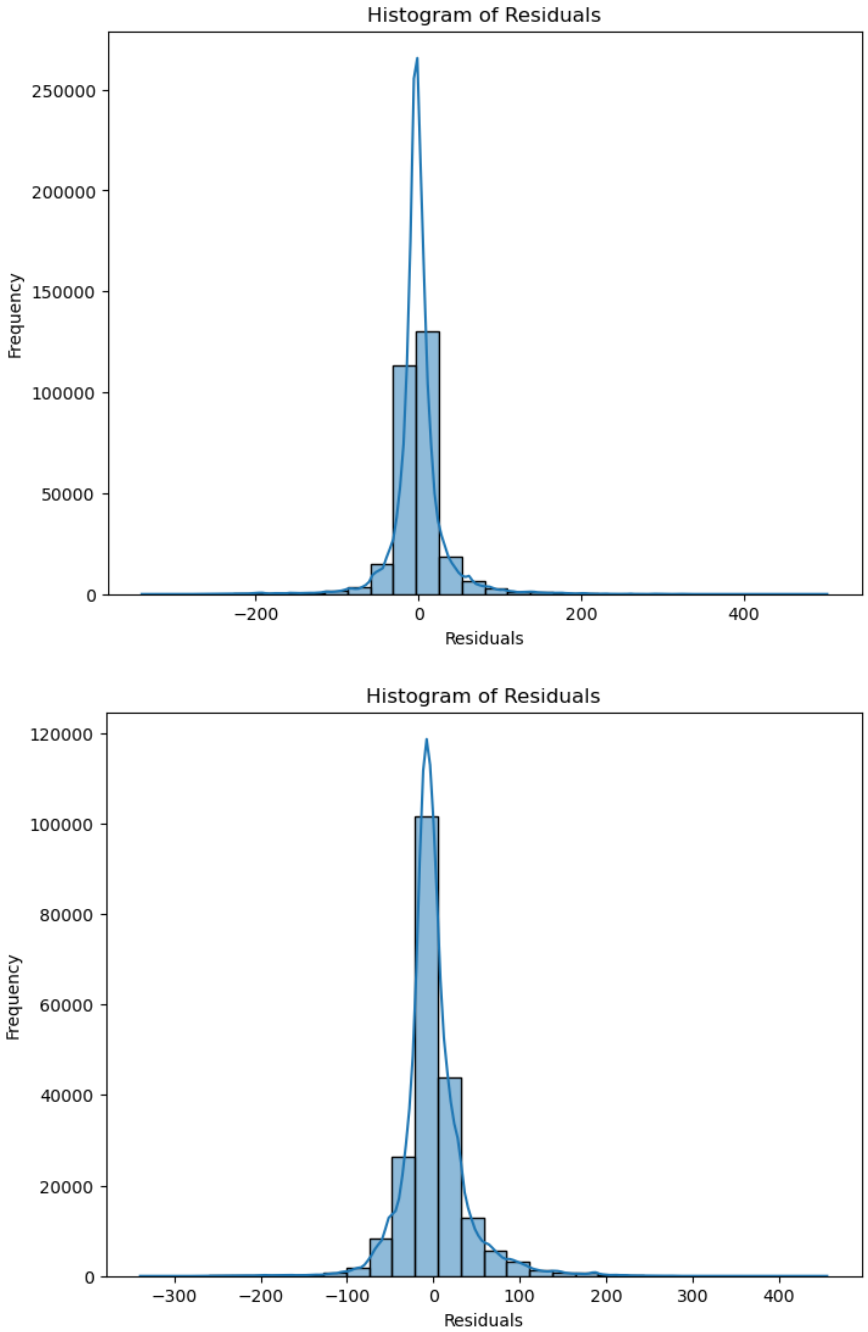

Figure.15: Histogram of Residuals (S and P, respectively)

3. Output of Model 2

a) Environment Group S

Results for Environment Group Code: S

OLS Regression Results

Dep. Variable:	value_sales	R-squared:	0.814
Model:	OLS	Adj. R-squared:	0.814
Method:	Least Squares	F-statistic:	2.647e+04
Date:	Wed, 06 Nov 2024	Prob (F-statistic):	0.00
Time:	04:55:45	Log-Likelihood:	-1.4517e+06
No. Observations:	295801	AIC:	2.904e+06
Df Residuals:	295751	BIC:	2.904e+06
Df Model:	49		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-27.9565	5.989	-4.668	0.000	-39.694	-16.219
value_cost	2.0643	0.002	1045.060	0.000	2.060	2.068
value_quantity	-0.1049	0.008	-12.708	0.000	-0.121	-0.089
business_area_code_920	15.8928	6.370	2.495	0.013	3.408	28.378
business_area_code_940	37.2018	6.295	5.910	0.000	24.864	49.540
business_area_code_945	29.0727	9.319	3.120	0.002	10.807	47.338
business_area_code_950	-28.5400	13.033	-2.190	0.029	-54.084	-2.996
business_area_code_960	15.1361	9.766	1.550	0.121	-4.004	34.276
business_area_code_970	24.8368	6.115	4.062	0.000	12.851	36.822
business_area_code_980	6.132e-14	3.15e-14	1.945	0.052	-4.7e-16	1.23e-13
business_area_code_985	-2.88e-14	6.89e-14	-0.418	0.676	-1.64e-13	1.06e-13
business_area_code_999	28.7478	7.411	3.879	0.000	14.223	43.272
business_area_code_COM	62.5879	5.999	10.433	0.000	50.830	74.346
business_area_code_DLT	6.2507	6.065	1.031	0.303	-5.637	18.138
business_area_code_EXL	5.136e-14	5.02e-14	1.023	0.306	-4.71e-14	1.5e-13
business_area_code_FLD	24.9649	5.996	4.163	0.000	13.212	36.717
business_area_code_HLB	-0.8054	6.091	-0.132	0.895	-12.744	11.133
business_area_code_IAE	4.371e-15	3.21e-14	0.136	0.892	-5.85e-14	6.72e-14
business_area_code_IAI	1.099e-13	6.31e-14	1.741	0.082	-1.38e-14	2.34e-13
business_area_code_LCP	3.2863	8.388	0.392	0.695	-13.155	19.727
business_area_code_LMP	30.1694	5.983	5.043	0.000	18.444	41.895
business_area_code_OTH	24.2169	5.983	4.048	0.000	12.491	35.943
business_area_code_PEN	39.0246	8.776	4.447	0.000	21.823	56.226
business_area_code_SAE	19.6236	10.140	1.935	0.053	-0.250	39.497
business_area_code_SUR	29.3833	5.998	4.899	0.000	17.628	41.138
business_area_code_TAL	42.1115	5.995	7.024	0.000	30.362	53.861
business_area_code_TRO	17.5052	6.869	2.548	0.011	4.042	30.968
business_area_code_URB	17.3210	6.562	2.640	0.008	4.459	30.183
calendar_month_2	0.8411	0.312	2.698	0.007	0.230	1.452
calendar_month_3	0.5134	0.312	1.647	0.099	-0.097	1.124
calendar_month_4	1.6410	0.313	5.238	0.000	1.027	2.255
calendar_month_5	0.7875	0.296	2.662	0.008	0.208	1.367
calendar_month_6	1.1887	0.305	3.902	0.000	0.592	1.786
calendar_month_7	0.1077	0.297	0.363	0.717	-0.474	0.690
calendar_month_8	0.2928	0.303	0.968	0.333	-0.300	0.886
calendar_month_9	0.3737	0.315	1.185	0.236	-0.245	0.992
calendar_month_10	0.1532	0.314	0.488	0.625	-0.462	0.768
calendar_month_11	1.5866	0.322	4.934	0.000	0.956	2.217
calendar_month_12	0.5668	0.352	1.611	0.107	-0.123	1.256
customer_district_code_210	1.1317	0.317	3.572	0.000	0.511	1.753
customer_district_code_300	-2.1762	0.176	-12.388	0.000	-2.520	-1.832
customer_district_code_310	-0.1542	0.413	-0.373	0.709	-0.965	0.656
customer_district_code_400	-2.0074	0.202	-9.941	0.000	-2.403	-1.612
customer_district_code_410	-1.1705	0.329	-3.560	0.000	-1.815	-0.526
customer_district_code_500	-1.6488	0.270	-6.108	0.000	-2.178	-1.120
customer_district_code_510	1.5223	0.533	2.856	0.004	0.478	2.567
customer_district_code_520	-49.6174	1.731	-28.670	0.000	-53.009	-46.225
customer_district_code_530	-2.3552	0.669	-3.518	0.000	-3.667	-1.043
customer_district_code_535	-4.9352	0.622	-7.929	0.000	-6.155	-3.715
customer_district_code_540	-5.8928	0.497	-11.862	0.000	-6.867	-4.919
customer_district_code_545	-25.2236	23.165	-1.089	0.276	-70.626	20.178
customer_district_code_600	1.2237	0.251	4.880	0.000	0.732	1.715
customer_district_code_710	77.9337	0.960	81.218	0.000	76.053	79.814
customer_district_code_720	-54.6678	0.653	-83.671	0.000	-55.948	-53.387

Omnibus:	83100.052	Durbin-Watson:	1.460
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3526961.379
Skew:	0.625	Prob(JB):	0.00
Kurtosis:	19.870	Cond. No.	1.33e+16

Figure.15: Model.2 (S)'s Output

b) Environment Group: P

Results for Environment Group Code: P

OLS Regression Results

Dep. Variable:	value_sales	R-squared:	0.822
Model:	OLS	Adj. R-squared:	0.822
Method:	Least Squares	F-statistic:	1.920e+04
Date:	Wed, 06 Nov 2024	Prob (F-statistic):	0.00
Time:	04:55:51	Log-Likelihood:	-1.0522e+06
No. Observations:	208344	AIC:	2.104e+06
Df Residuals:	208293	BIC:	2.105e+06
Df Model:	50		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-31.7836	6.395	-4.970	0.000	-44.318	-19.249
value_cost	1.7693	0.002	839.788	0.000	1.765	1.773
value_quantity	0.4816	0.018	26.263	0.000	0.446	0.518
business_area_code_920	72.8919	27.455	2.655	0.008	19.081	126.703
business_area_code_940	53.1651	8.110	6.555	0.000	37.270	69.060
business_area_code_945	40.6400	38.298	1.061	0.289	-34.424	115.704
business_area_code_950	40.5840	8.691	4.669	0.000	23.549	57.619
business_area_code_960	61.9287	9.165	6.757	0.000	43.966	79.891
business_area_code_970	7.282e-13	4.76e-13	1.529	0.126	-2.05e-13	1.66e-12
business_area_code_980	54.2059	8.742	6.201	0.000	37.073	71.339
business_area_code_985	57.9900	38.298	1.514	0.130	-17.073	133.053
business_area_code_999	369.5938	9.783	37.780	0.000	350.420	388.768
business_area_code_COM	47.0471	6.390	7.362	0.000	34.522	59.572
business_area_code_DLT	32.8108	6.395	5.131	0.000	20.277	45.345
business_area_code_EXL	44.9690	6.691	6.721	0.000	31.856	58.082
business_area_code_FLD	47.7793	6.392	7.475	0.000	35.252	60.307
business_area_code_HLB	34.3829	6.428	5.349	0.000	21.784	46.982
business_area_code_IAE	8.859e-13	2.58e-13	3.440	0.001	3.81e-13	1.39e-12
business_area_code_IAI	-8.537e-13	2.26e-13	-3.786	0.000	-1.3e-12	-4.12e-13
business_area_code_LCP	35.4252	6.594	5.372	0.000	22.500	48.350
business_area_code_LMP	41.1550	6.394	6.437	0.000	28.623	53.686
business_area_code_OTH	38.4311	6.398	6.006	0.000	25.890	50.972
business_area_code_PEN	42.8666	7.700	5.567	0.000	27.775	57.958
business_area_code_RWY	-5.662e-13	4.29e-13	-1.321	0.187	-1.41e-12	2.74e-13
business_area_code_SAE	49.8774	6.396	7.798	0.000	37.341	62.413
business_area_code_SUR	46.4619	6.386	7.275	0.000	33.945	58.979
business_area_code_TAL	43.5707	6.423	6.784	0.000	30.982	56.159
business_area_code_TRO	30.5929	6.399	4.781	0.000	18.052	43.134
business_area_code_URB	41.9381	6.535	6.417	0.000	29.129	54.747
calendar_month_2	0.5027	0.424	1.185	0.236	-0.329	1.334
calendar_month_3	1.4607	0.430	3.400	0.001	0.619	2.303
calendar_month_4	2.5807	0.428	6.027	0.000	1.741	3.420
calendar_month_5	2.8511	0.414	6.892	0.000	2.040	3.662
calendar_month_6	1.3626	0.427	3.190	0.001	0.525	2.200
calendar_month_7	0.0599	0.413	0.145	0.885	-0.751	0.870
calendar_month_8	-0.4157	0.418	-0.995	0.320	-1.234	0.403
calendar_month_9	0.8754	0.433	2.020	0.043	0.026	1.725
calendar_month_10	0.9359	0.433	2.159	0.031	0.086	1.785
calendar_month_11	2.4052	0.439	5.478	0.000	1.545	3.266
calendar_month_12	1.4824	0.482	3.076	0.002	0.538	2.427
customer_district_code_210	-1.6634	0.473	-3.516	0.000	-2.591	-0.736
customer_district_code_300	-3.3834	0.272	-12.422	0.000	-3.917	-2.850
customer_district_code_310	0.3483	0.640	0.544	0.586	-0.906	1.603
customer_district_code_400	-2.5391	0.293	-8.666	0.000	-3.113	-1.965
customer_district_code_410	1.2357	0.437	2.828	0.005	0.379	2.092
customer_district_code_500	-5.4021	0.343	-15.772	0.000	-6.073	-4.731
customer_district_code_510	4.3483	0.641	6.778	0.000	3.091	5.606
customer_district_code_520	-30.3103	2.491	-12.165	0.000	-35.194	-25.427
customer_district_code_530	2.4627	0.807	3.053	0.002	0.882	4.044
customer_district_code_535	-5.4236	0.903	-6.003	0.000	-7.194	-3.653
customer_district_code_540	-3.4549	0.631	-5.477	0.000	-4.691	-2.218
customer_district_code_545	139.3492	15.419	9.037	0.000	109.128	169.571
customer_district_code_600	-1.9432	0.340	-5.715	0.000	-2.610	-1.277
customer_district_code_710	-37.3113	1.206	-30.944	0.000	-39.675	-34.948
customer_district_code_720	-50.2499	0.348	-144.533	0.000	-50.931	-49.568

Figure.16: Model.2(P)'s Output

Overall, the "*P*" model achieved the highest performance with an R-squared value of 82.2%, followed closely by model "*S" (81.4%).* Accordingly, the models for environmental groups P and S demonstrated a positive significant relationship with both *business_area_codes* and *calendar_month*, while showing a negative relationship with *customer_district_code*.

Both models also presented insignificant relationships with *business_code_985*, customer district 310, and during the months of *July* and *August*. Additionally, remarkable high coefficients were observed in certain areas, such as *business_area_code_999* and *business_area_code_COMP*, compared to other factors, indicating that the effect of these business areas varies significantly.

VI. Prediction Models

1. Optimal Model for Sales Prediction

The decision to use both Multiple Linear Regression (MLR) and Random Forest (RF) models for prediction is based on their complementary strengths. MLR provides transparency and clear relationships between sales and predictors like **cost, quantity, and profit margin**, which are significant to sales performance (**Model.1**). RF, on the other hand, captures complex, non-linear relationships and helps identify the most important factors influencing sales. The feature selection from RF also highlighted **cost, quantity, and profit margin** as key predictors (Figure.17), justifying their inclusion in the MLR model.

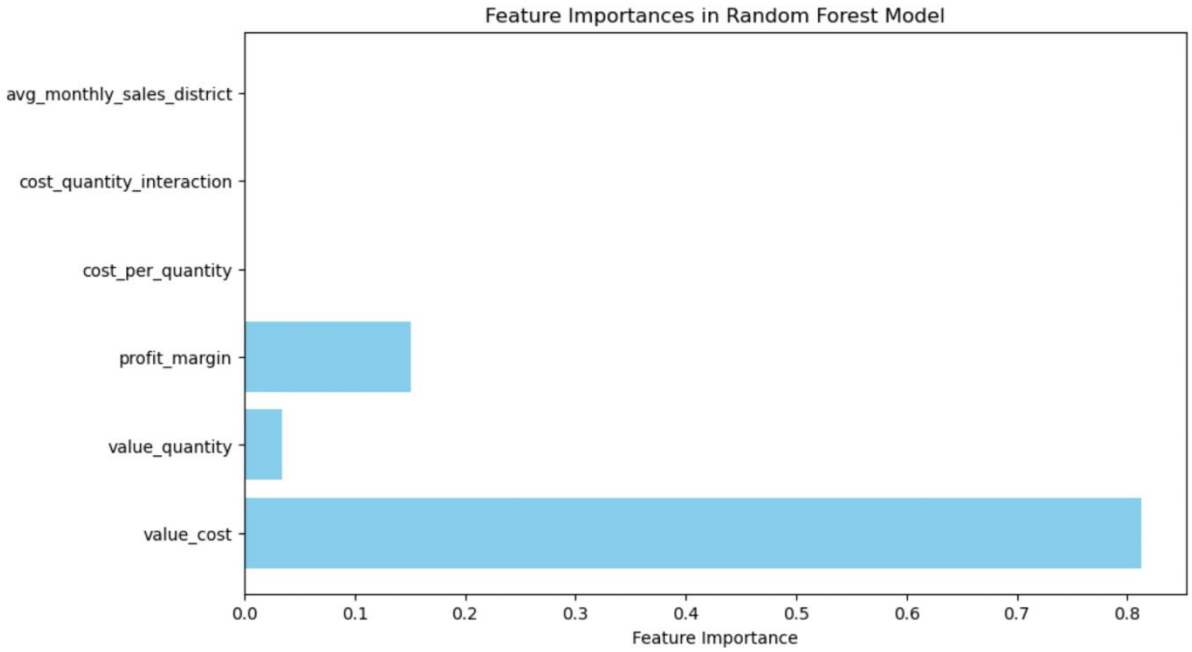

Figure.17: Feature Importances in Random Forest Model

2. Model Performance Comparisons

In comparing the performance of the multiple linear regression (MLR) and random forest (RF) models, the results show a clear distinction in their accuracy. Also, The MAE for the MLR and RF models is 17.61 and 1.1, showing the RF model's superior accuracy, with predictions deviating by only **$1.10 on average**, compared to **$17.61 for MLR** from actual value sales. Similarly, the RMSE for MLR and RF's are 31.44 and 9.29, respectively, showing the RF model's superior accuracy, with the average magnitude of larger prediction errors are **$31.44 and $9.29**. Using both *MAE* and *RMSE* provides a well-rounded evaluation, highlighting both overall accuracy and the model's ability to handle outliers. This suggests that while the MLR model provides reasonable predictions, the RF model demonstrates superior overall performance, particularly in handling large errors. Given the considerable advantage in both MAE and RMSE, the RF model is the preferable choice for this prediction task.

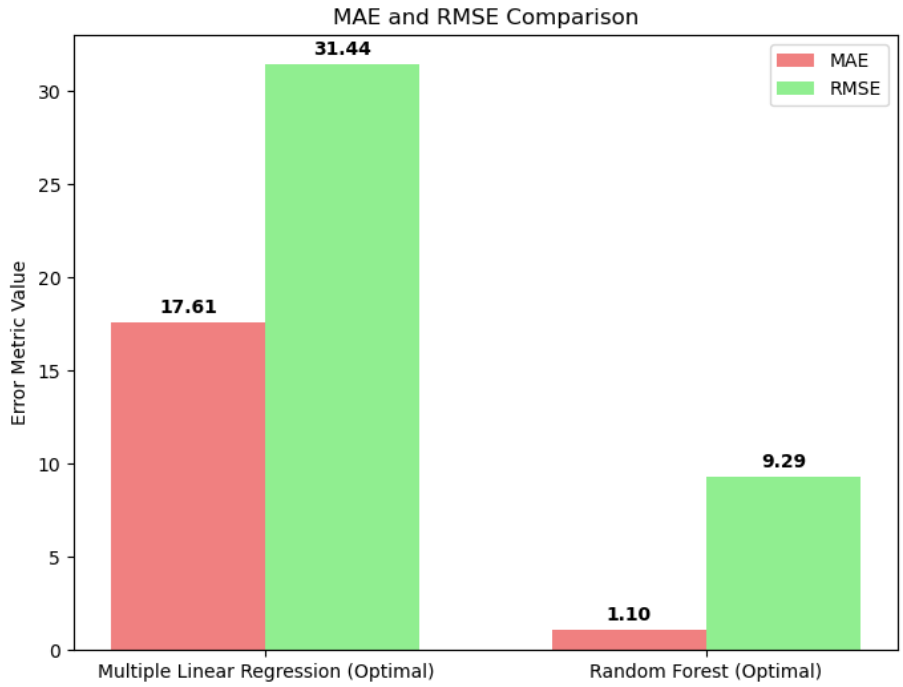

Figure.18: MAE and RMSE comparison between MLR and RF models

VII. Higher Likelihood of Losing Customers

This section employs LR and RF models to predict the likelihood of customer churn, assigning a value of 1 and 0 for churn and non-churn, respectively. According to Figure.19, the feature selection process from the RF model has identified key drivers of customer churn at LuminaTech, including **(1) "Time Gap" (Delivery Delays)**, which is the strongest predictor of churn, as delays in order fulfillment significantly increase the risk of customer attrition, **(2) "Monetary Value" (Total Spend)**, which reveals the vulnerability of high-value customers, where a decline in spending may suggest potential churn, **(3) "Purchase Frequency**", where

a decrease indicates disengagement and further heightens the churn risk, **(4) "Refund Patterns",** which play a significant role, as higher amounts and frequencies of refunds suggest customer dissatisfaction, and **(5) Recency** as the Figure.20 indicates that **90 Days Since Last Purchase** is a stronger predictor, with customers who have longer gaps since their last purchase being more likely to churn. However, the key predictor of recency is defined as 6 months since the last order, as lightning products typically have a longer lifespan of several months to years (Morgan, 2012).

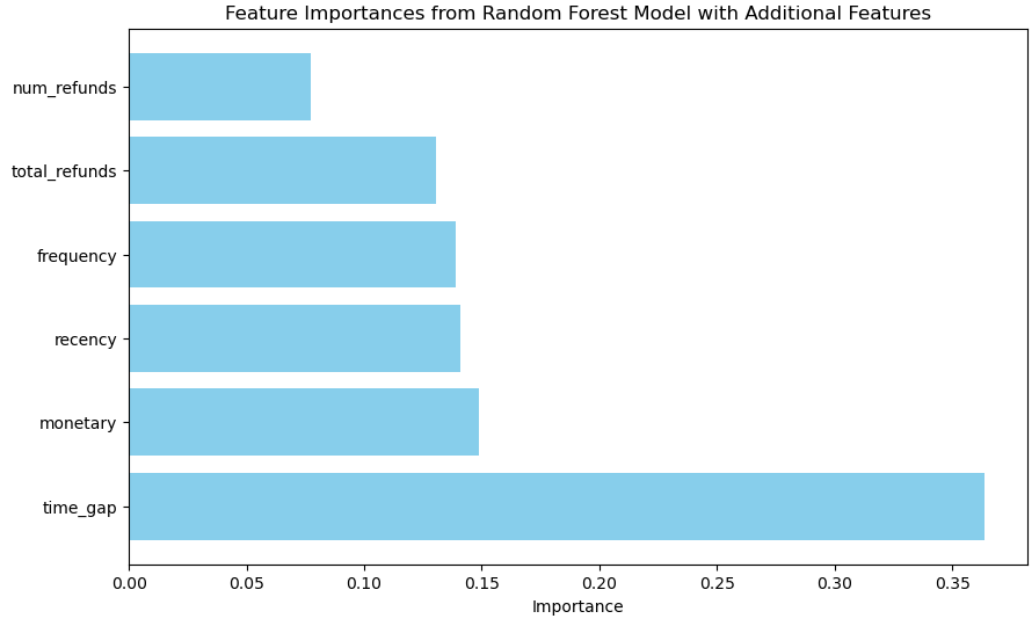

Figure.19: Feature Importances from Random Forest Model with Additional Features

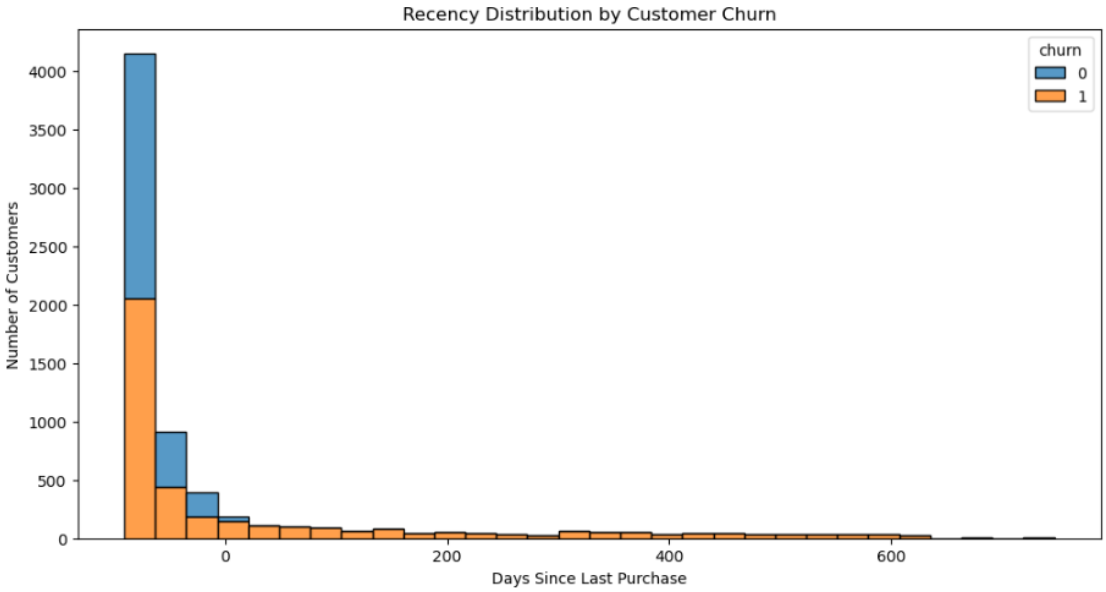

Figure.20: Recency Distribution by Customer Churn

3. MLR and RF model performance comparison

a) Logistic Regression Model (LR)

The LR model shows moderate performance for churned customers (precision 0.77, recall 1.00) but fails completely for non-churned customers (precision and recall both 0.00). Despite an overall accuracy of 0.77, the model is heavily biased towards churned customers, as shown by the poor macro and weighted averages (Figure.21)

Logistic Regression Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	65253
1	0.77	1.00	0.87	212679
accuracy			0.77	277932
macro avg	0.38	0.50	0.43	277932
weighted avg	0.59	0.77	0.66	277932

Figure.21: Logistic Regression Report

b) Random Forest Model (RF)

RF model (Figure.22) performs well for churned customers, with a precision of 0.78 and recall of 0.98. However, it faces challenges in predicting non-churned customers, with precision of 0.65 and recall of only 0.12. The overall accuracy of the model is 0.78, but the significant imbalance in predictions highlights the model's bias towards churned customers.

Random Forest Report:

	precision	recall	f1-score	support
0	0.65	0.12	0.21	65253
1	0.78	0.98	0.87	212679
accuracy			0.78	277932
macro avg	0.72	0.55	0.54	277932
weighted avg	0.75	0.78	0.71	277932

Figure.22: Random Forest Report

4. Confusion matrix:

a) Logistic Regression Model (LR)

The confusion matrix results indicate that the Logistic Regression model performs well in predicting churned customers, with 212,679 true positives (TP), meaning it accurately identifies customers who are likely to churn. However, the model struggles with non-churned customers, as it predicts 65,253 non-churned customers as churned (false positives, FP), leading to resource wastage on retention efforts for customers who are unlikely to leave. There are no true negatives (TN) or false negatives (FN), meaning the model does not correctly identify any non-churned customers, nor does it miss any churned customers. This highlights

a strong bias towards predicting churned customers, but a need for improvement in detecting non-churned customers.

b) Random Forest Model (RF)

The confusion matrix for the RF model shows it performs well at identifying churned customers, with 208,327 (TP), similar to the LR model's 212,679 TP. However, the RF model misclassifies 57,280 non-churned customers as churned (FP), which is higher than LR's 65,253 FP. The RF model has fewer false negatives (4,352), **indicating better performance in identifying churned customers**. However, both models struggle with class imbalance, leading to biases towards churned customers and misclassifications for non-churned ones. While both models are effective for predicting churned customers, the RF model's higher false positive rate makes it less reliable for predicting non-churned customers. Both would benefit from techniques to address class imbalance.

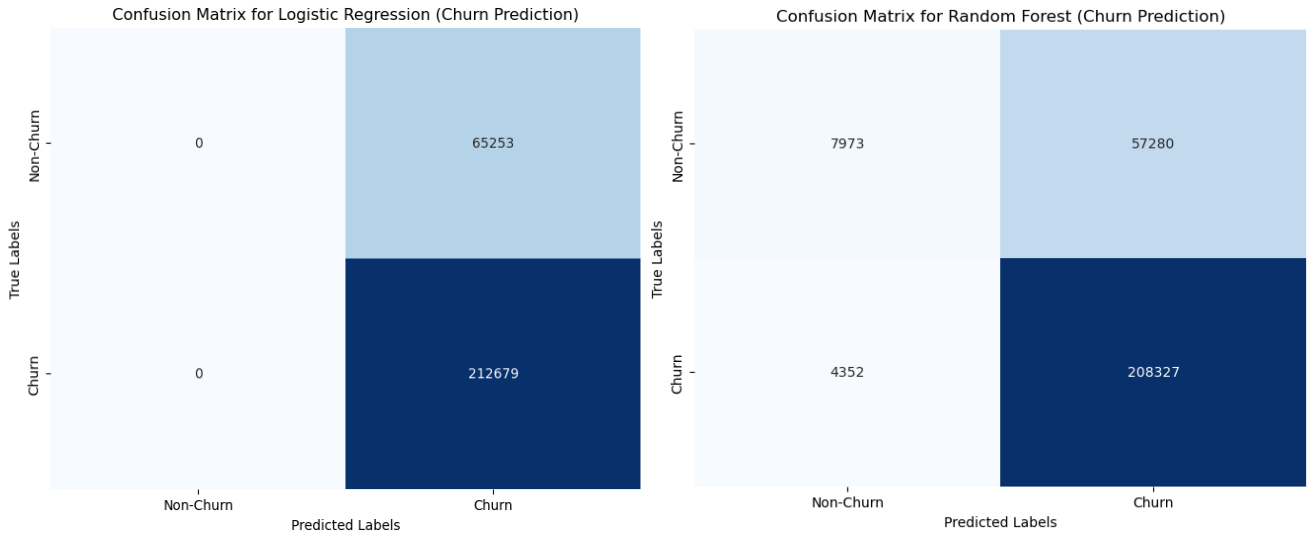

Figure.23: Confusion Matrix for LR and RF models, respectively

5. Imbalance handling

Upon further investigation of the dataset, a significant class imbalance between churned and non-churned customers was identified based on transactions (Figure.24).

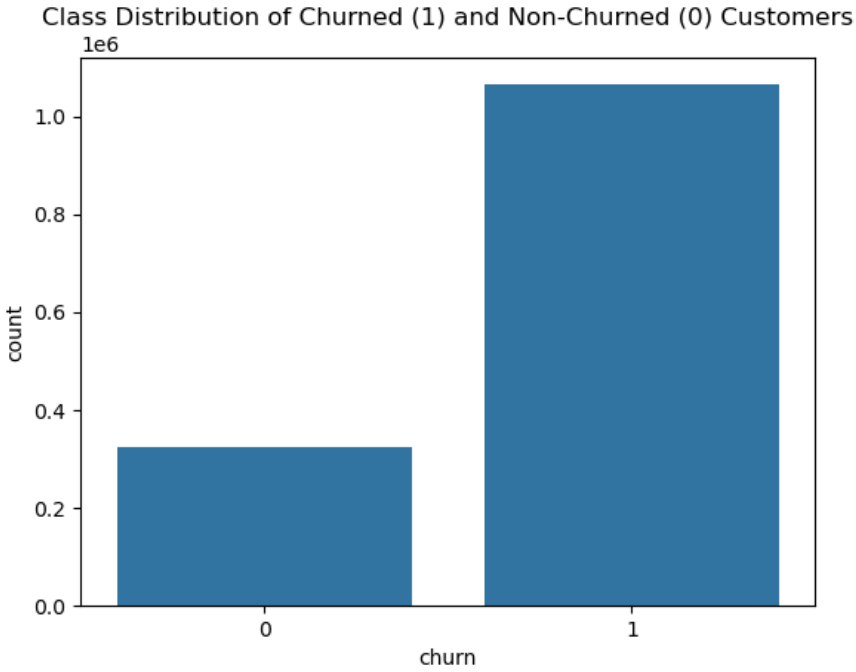

Figure.24: Class Distribution for Churn and Non- Churn

6. Classification (Before and After Undersampling)

Given this imbalance, initial classification results showed poor performance for non-churned customers (precision 0.24, recall 0.83) and decent results for churned customers (precision 0.81, recall 0.23). The overall accuracy of 0.37 highlighted the class imbalance, with the model favouring non-churned customers. The recall for the churned class was 0.23, meaning that the model was only capturing 23% of the actual churned cases. This low recall suggests that the model was missing a substantial portion of customers who would churn, limiting its usefulness in scenarios where identifying potential churners is crucial. Moreover, the F1-score, which balances precision and recall, was 0.35 for the churned class, reflecting the model's inability to balance true positive predictions with false positives adequately.

Classification Report:

	precision	recall	f1-score	support
Non-Churned	0.24	0.83	0.38	64758
Churned	0.81	0.23	0.35	213174
accuracy			0.37	277932
macro avg	0.53	0.53	0.37	277932
weighted avg	0.68	0.37	0.36	277932

Figure.25: Classification before Undersampling

To address the imbalance, undersampling was applied to balance the classes. After undersampling, the model's accuracy improved to 0.72. However, the performance for non-churned customers remained poor (precision 0.25 and recall 0.1), while the performance for **churned customers improved significantly** (precision 0.77 and recall 0.91). Recall for churned customers improving from 0.23 to 0.91 means the model now **correctly identifies 91% of the actual churned customers**, compared to only **23% before undersampling**. This is crucial because high recall ensures that fewer churned customers are missed, which is important for a business that aims to take action on customer at risk of leaving. Despite the improvement, the model remained biased towards churned customers due to the imbalance.

Classification Report After Undersampling:

	precision	recall	f1-score	support
Non-Churned	0.25	0.10	0.14	64758
Churned	0.77	0.91	0.83	213174
accuracy			0.72	277932
macro avg	0.51	0.50	0.49	277932
weighted avg	0.65	0.72	0.67	277932

Figure.26: Classification After Undersampling

The post-undersampling model now excels at identifying almost all potential churners (91% recall for churned class), making it highly effective for pre-emptive retention strategies. The increased F1-score for the churned class means that the model balances accuracy and reliability in identifying churners, which is essential for prioritizing retention efforts effectively.

VIII. Key Insights and Recommendations

The company is facing several key challenges that are impacting its performance. First, its product strategy is not optimized, with an over-reliance on "make-to-order" products, which involve longer lead times, higher upfront costs, and complexity, despite placing less burden on inventory costs (Doug Bulla, 2022). However, the delivery time may influence current customer decisions. This strategy reduces market share in the "make-to-stock" product segment and potentially increases churn, especially when delivery times are inconsistent across customer groups. The high return/refund rates are strongly correlated with churn, highlighting the need to address these issues.

Secondly, seasonality affects revenue, with notable low sales periods in July and August and during the summer months (particularly December and January), while profit margins peak in September and November. This underscores the need for better demand forecasting, sales strategy, and inventory management during peak seasons. Additionally, the company's long-

lasting products result in unrealistic expectations of repeat purchases in the short and medium term (Morgan, 2012). There is also a lack of diversification in the product portfolio, which may be contributing to a loss of customers, as the market demand is not fully addressed. The high return rates, reaching 50%-80% for some customers, further indicate potential product issues that require further investigation. Lastly, although non-project orders are generally delivered on time, project orders face delays, suggesting that improvements in production or supply chain processes are necessary.

Therefore, the company should emphasise non-project orders related to high-margin environmental product groups "P" and "I" and focus marketing efforts on high-profit districts, such as 530, 535, and 520, as well as business_area_code_999, COMP and SUR. At the same time, the company should address challenges in underperforming districts, such as 545, 300, and 400. Inventory management should be optimized with the sales prediction to ensure sufficient stock during peak profit months while adjusting sales strategies for lower demand during off-peak months, applying a seasonal lightning concept strategy (Veeqo, n.d.; Marshall, 2023). Simultaneously, purchasing through subscriptions, loyalty rewards, or personalized promotions are also recommended (McKinsey, 2024). Understanding the product-mix demand and expanding the product range by districts will help diversify and maintain offerings, reducing reliance on custom products.

Finally, to mitigate the high return rates, the company should investigate the root causes and improve product quality, descriptions, and customer service through feedback loops (Momaya, 2022). Finally, optimizing logistics and delivery processes, particularly for project orders, will help reduce delays and improve customer satisfaction (Chiarini and Douglas, 2015). Strategic initiatives should include targeted marketing campaigns for specified regions and products, loyalty programs to drive revenue during peak seasons, and a focus on high-profit products and regions to maximise overall business performance.

REFERENCE

Aboutspacelightning. (2017). *Project Lights Made To Order | About Space Lighting*. About Space. Https://Www.Aboutspace.Net.Au/Collections/Project-Lights?Srsltid=Afmbooqm-Mbnfn5vtl282scvaqiv7nb2seurxkggq1rpknv6jsnvyhua

Bulla, D. (2022). *Https://Mcaconnect.Com/Resources/Its-Time-To-Move-To-Make-To-Order/*. Mcaconnect. Https://Mcaconnect.Com/Resources/Its-Time-To-Move-To-Make-To-Order

Fernandes, T., & Meena Rambocas. (2013). *Evaluating The Impact Of Customer Demographical Characteristics On Relationship Outcomes*. Researchgate; Unknown. Https://Www.Researchgate.Net/Publication/263925425_Evaluating_The_Impact_Of_Customer_Demographical_Characteristics_On_Relationship_Outcomes

Jinal Momaya, & Muley, K. (2022). Customer Feedback System & Businesses. *Researchgate*. Https://Doi.Org/10.55041/Ijsrem15314

Marshall, B. (2023, March). *Seasonal Lighting*. Se Lighting. Https://Selighting.Com.Au/Blogs/News/Seasonal-Lighting

Mckinsey. (2024, April 3). *Members Only: Delivering Greater Value Through Loyalty And Pricing | Mckinsey*. Www.Mckinsey.Com. Https://Www.Mckinsey.Com/Capabilities/Growth-Marketing-And-Sales/Our-Insights/Members-Only-Delivering-Greater-Value-Through-Loyalty-And-Pricing

Morgan, F. (2012). *Gauging The Lifetime Of An Led*. Maser.Com. Https://Maser.Com.Au/Wp-Content/Uploads/2017/03/Digital-Lumens-Gauging-Led-Lifetime.Pdf

Özsu, M., Ganti, V., & Das Sarma, A. (2013). *Synthesis Lectures On Data Management Data Cleaning A Practical Perspective Mor Gan Cl Aypool Publishers & Synthesis Lectures On Data Management Data Cleaning A Practical Perspective Mor Gan Cl Aypool Publishers & Synthesis Lectures On Data Management Data Cleaning A Practical Perspective*. Https://Odbms.Org/Wp-Content/Uploads/2014/03/Data-Cleaning.Pdf

Pohl, M., Staegemann, D. G., & Turowski, K. (2022). The Performance Benefit Of Data Analytics Applications. *Procedia Computer Science*, *201*, 679–683. Sciencedirect. Https://Doi.Org/10.1016/J.Procs.2022.03.090

Retaileconomic. (2023). *Peak Season Report 2023 | Retail Economics*. Retaileconomics.Co.Uk. Https://Www.Retaileconomics.Co.Uk/Retail-Insights/Thought-Leadership-Reports/Peak-Trading-Season-Report-2023

Sustainability Method. (N.D.). *Barplots, Histograms And Boxplots*. Sustainability Method. Https://Sustainabilitymethods.Org/Index.Php/Barplots,_Histograms_And_Boxplots

Veeqo. (N.D.). The Complete Guide To Inventory Management 1. In *Assets.Ctfassets*. Https://Assets.Ctfassets.Net/Hfb264dqso7g/3kmzofgmm9paadvxhtelxj/Cf96f49029129793bd82811b43406ac7/Inventory-Management-Pdf.Pdf

APPENDIX

Appendix.1: Top 10 profit margin of environmental_group_code in business_area_code

	environment_group_code	business_area_code	avg_profit_margin	total_sales	total_quantity
37	P	999	0.978191	8.898780e+03	26.0
21	I	DLT	0.848487	3.950000e+02	1.0
2	C	970	0.805471	8.548320e+03	2688.0
0	C	920	0.743580	1.154700e+03	23.0
91	Z	SAE	0.700218	1.415267e+03	8.0
15	C	TAL	0.688849	9.511673e+05	44471.0
22	I	IAE	0.686064	5.162591e+03	140.0
10	C	LCP	0.671844	2.962952e+05	26109.0
52	R	985	0.645600	5.300000e+02	2.0
11	C	LMP	0.632515	1.025611e+07	3345656.0

Appendix.2: Linearity Assumption Checking for Model.2 (S)

value_sales	1.000000
value_cost	0.901357
value_quantity	0.152441
business_area_code_920	0.011102
business_area_code_940	0.017219
business_area_code_945	0.004149
business_area_code_950	0.004590
business_area_code_960	0.007571
business_area_code_970	-0.001587
business_area_code_980	0.011980
business_area_code_985	0.007611
business_area_code_999	0.005158
business_area_code_COM	0.089699
business_area_code_DLT	0.016055
business_area_code_EXL	0.013455
business_area_code_FLD	0.083385
business_area_code_HLB	0.027896
business_area_code_IAE	0.004609
business_area_code_IAI	-0.001164
business_area_code_LCP	0.005063
business_area_code_LMP	-0.238656
business_area_code_OTH	0.002403
business_area_code_PEN	-0.002574
business_area_code_RWY	0.077777
business_area_code_SAE	0.089466
business_area_code_SUR	0.110229
business_area_code_TAL	0.011463
business_area_code_TRO	0.044379
business_area_code_URB	0.042831
calendar_month_2	0.001773
calendar_month_3	0.000866
calendar_month_4	0.003860
calendar_month_5	0.009288
calendar_month_6	0.001856
calendar_month_7	-0.004187
calendar_month_8	-0.000410
calendar_month_9	-0.005209
calendar_month_10	-0.005831
calendar_month_11	0.008302
calendar_month_12	-0.004156
customer_district_code_210	-0.008913
customer_district_code_300	0.006353
customer_district_code_310	0.009534
customer_district_code_400	-0.022147
customer_district_code_410	0.022885
customer_district_code_500	-0.042018
customer_district_code_510	0.030196
customer_district_code_520	-0.009690
customer_district_code_530	0.011119
customer_district_code_535	0.004558

```
business_area_code_SUR        0.110229
business_area_code_TAL        0.011463
business_area_code_TRO        0.044379
business_area_code_URB        0.042831
calendar_month_2              0.001773
calendar_month_3              0.000866
calendar_month_4              0.003860
calendar_month_5              0.009288
calendar_month_6              0.001856
calendar_month_7             -0.004187
calendar_month_8             -0.000410
calendar_month_9             -0.005209
calendar_month_10            -0.005831
calendar_month_11             0.008302
calendar_month_12            -0.004156
customer_district_code_210   -0.008913
customer_district_code_300    0.006353
customer_district_code_310    0.009534
customer_district_code_400   -0.022147
customer_district_code_410    0.022885
customer_district_code_500   -0.042018
customer_district_code_510    0.030196
customer_district_code_520   -0.009690
customer_district_code_530    0.011119
customer_district_code_535    0.004558
customer_district_code_540    0.007687
customer_district_code_545   -0.001670
customer_district_code_600    0.052625
customer_district_code_710   -0.004474
customer_district_code_720   -0.010824
Name: value_sales, dtype: float64
```

Appendix.3: Linearity Assumption Checking for Model.2 (P)

```
value_sales                    1.000000
value_cost                     0.901357
value_quantity                 0.152441
business_area_code_920         0.011102
business_area_code_940         0.017219
business_area_code_945         0.004149
business_area_code_950         0.004590
business_area_code_960         0.007571
business_area_code_970        -0.001587
business_area_code_980         0.011980
business_area_code_985         0.007611
business_area_code_999         0.005158
business_area_code_COM         0.089699
business_area_code_DLT         0.016055
business_area_code_EXL         0.013455
business_area_code_FLD         0.083385
business_area_code_HLB         0.027896
business_area_code_IAE         0.004609
business_area_code_IAI        -0.001164
business_area_code_LCP         0.005063
business_area_code_LMP        -0.238656
business_area_code_OTH         0.002403
business_area_code_PEN        -0.002574
business_area_code_RWY         0.077777
business_area_code_SAE         0.089466
business_area_code_SUR         0.110229
business_area_code_TAL         0.011463
business_area_code_TRO         0.044379
business_area_code_URB         0.042831
calendar_month_2               0.001773
calendar_month_3               0.000866
calendar_month_4               0.003860
calendar_month_5               0.009288
calendar_month_6               0.001856
calendar_month_7              -0.004187
calendar_month_8              -0.000410
calendar_month_9              -0.005209
calendar_month_10             -0.005831
calendar_month_11              0.008302
calendar_month_12             -0.004156
customer_district_code_210    -0.008913
customer_district_code_300     0.006353
customer_district_code_310     0.009534
customer_district_code_400    -0.022147
customer_district_code_410     0.022885
customer_district_code_500    -0.042018
customer_district_code_510     0.030196
customer_district_code_520    -0.009690
customer_district_code_530     0.011119
customer_district_code_535     0.004558
customer_district_code_540     0.007687
customer_district_code_545    -0.001670
customer_district_code_600     0.052625
customer_district_code_710    -0.004474
customer_district_code_720    -0.010824
Name: value_sales, dtype: float64
```

Appendix.4: Foreign Exchange Rates (Conversions of Curerncies)